

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-263514

(43)Date of publication of application : 11.10.1996

(51)Int.Cl.

G06F 17/30

G06F 7/24

(21)Application number : 07-068160

(71)Applicant : MITSUBISHI ELECTRIC CORP

(22)Date of filing : 27.03.1995

(72)Inventor : ARITA HIDEKAZU

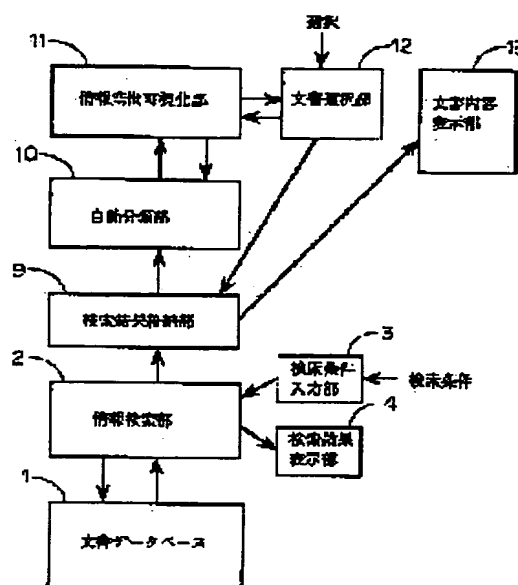
YASUI TERUMASA

TSUDAKA SHINICHIRO

(54) METHOD FOR AUTOMATIC CLASSIFICATION OF DOCUMENT, METHOD FOR VISUALIZATION OF INFORMATION SPACE, AND INFORMATION RETRIEVAL SYSTEM**(57)Abstract:**

PURPOSE: To obtain the information retrieval system which can easily obtain a retrieval key word and functions as a bottom-up type sending support system.

CONSTITUTION: This system is provided with a retrieval result storage part 9 in which a set of documents retrieved from a document data base 1 by an information retrieval part 2 under retrieval conditions from a retrieval condition input part 3 is stored, an automatic classification part 10 which automatically classifies the documents in the document set, an information space visualization part 11 which visualizes the information space of the automatically classified document set, a document selection part 12 which calls an automatically classified document by specifying the visualized two-dimensional position and selects one of plural documents when there are plural documents, and a document content display part 13 which takes the contents of the selected document out of the retrieval storage part 9 and displays them.

**LEGAL STATUS**

[Date of request for examination]

01.07.1999

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

3385297

[Date of registration]

27.12.2002

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-263514

(43) 公開日 平成8年(1996)10月11日

(51) IntCl. ⁸	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 17/30 7/24		9194-5L 9194-5L 9194-5L	G 0 6 F 15/403 7/24 15/403	3 5 0 Z L 3 4 0 B 3 5 0 C

審査請求 未請求 請求項の数 6 O L (全 18 頁)

(21) 出願番号 特願平7-68160

(22) 出願日 平成7年(1995)3月27日

(71) 出願人 000006013

三菱電機株式会社

東京都千代田区丸の内二丁目2番3号

(72) 発明者 有田 英一

尼崎市塚口本町八丁目1番1号 三菱電機株式会社中央研究所内

(72) 発明者 安井 照昌

尼崎市塚口本町八丁目1番1号 三菱電機株式会社中央研究所内

(72) 発明者 津高 新一郎

尼崎市塚口本町八丁目1番1号 三菱電機株式会社中央研究所内

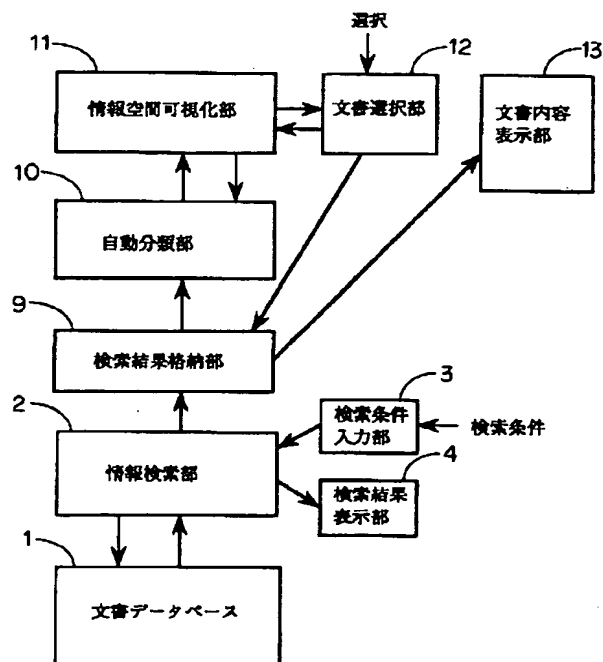
(74) 代理人 弁理士 田澤 博昭 (外2名)

(54) 【発明の名称】 文書の自動分類方法、および情報空間の可視化方法、ならびに情報検索システム

(57) 【要約】

【目的】 検索キーワードが容易に得られ、ボトムアップ型の発注支援システムとして機能する情報検索システムを得る。

【構成】 検索条件入力部3からの検索条件に従って情報検索部2が文書データベース1より検索した文書集合を格納する検索結果格納部9、その文書集合を対象に文書を自動分類する自動分類部10、自動分類された文書集合の情報空間を視覚化する情報空間可視化部11、視覚化された2次元の位置を指定することにより自動分類された文書の呼び出しを行い、それが複数ある場合にはその中より1つを選択する文書選択部12、および選択された文書の内容を検索結果格納部より取り出して表示する文書内容表示部13を設けたもの。



【特許請求の範囲】

【請求項 1】 文書集合が与えられた時、その文書集合中の各文書を、その内容に応じて自動的に分類する文書の自動分類方法において、前記文書集合に含まれる、単語や句、節などの意味のある文字列による語句の中から、一定の条件で選んだ語句を構成要素として、文書集合中のある文書を、それが含む語句に対応する値をその語句の出現頻度をもとに定めた語句ベクトルとして表現し、前記文書の語句ベクトルと 2 次元に配置されたセルに対応する語句ベクトルの距離を計算して、その距離が最小のものをその文書が所属する仮のセルとし、当該セルの語句ベクトルの要素の値をその文書の語句ベクトルの要素の値に近付けるとともに、そのセルの近傍のセルの語句ベクトルの要素の値も、その文書の語句ベクトルに対する近傍の度合いに応じて減じて近付けることを、前記文書集合に含まれる文書について一定回数、もしくは収束するまで実行し、その後、各セルの語句ベクトルと文書の語句ベクトルとの距離を計算して、その距離が最小のセルをその文書が所属する本来のセルとして、同じセルに所属する文書を内容が類似した文書のクラスと判断することを特徴とする文書の自動分類方法。

【請求項 2】 前記文書が所属する仮のセルが決まった後に、近傍のセルの語句ベクトルを文書の語句ベクトルに近付ける際、前記文書集合での語句の出現の割合に依存して、前記近傍のセルの語句ベクトルを文書の語句ベクトルに近付ける割合を変化させることにより、文書のクラスターの階層性と文書クラスターの所属するセルの領域の包含関係を対応させることを特徴とする請求項 1 に記載の文書の自動分類方法。

【請求項 3】 請求項 1 または 2 に記載の文書の自動分類方法によって分類された文書集合の情報空間を可視化する情報空間の可視化方法において、セルに対応する語句ベクトルのうちの、値の大きなものから一定数、もしくは所定の閾値以上の値を持つものを、そのセルを代表する語句として、その文字列の表示を行うことを特徴とする情報空間の可視化方法。

【請求項 4】 隣り合ったセルの境界線の属性を、セルに対応する語句ベクトルの距離に応じて変化させて表示することを特徴とする請求項 3 に記載の情報空間の可視化方法。

【請求項 5】 複数の文書が格納された文書データベースと、前記文書データベースより検索条件に該当する文書を検索抽出する情報検索部と、前記情報検索部に検索条件を入力する検索条件入力部と、前記情報検索部による検索結果を出力表示する検索結果表示部とを有する情報検索システムにおいて、前記情報検索部にて検索された文書集合を格納する検索結果格納部と、前記検索結果格納部に格納された文書集合を対象に、請求項 1 または 2 に記載した文書の自動分類方法によって文書の自動分類を行う自動分類部と、前記自動分類部にて分類された

文書集合の情報空間を、請求項 3 または 4 に記載した情報空間の可視化方法によって視覚化する情報空間可視化部と、前記情報空間可視化部にて視覚化された 2 次元の位置を指定することによって前記自動分類された文書の選択を行い、それが複数ある場合にはその中の所定数の文書を選択する文書選択部と、前記文書選択部にて選択された文書の内容を前記検索結果格納部より取り出して表示する文書内容表示部を備えたことを特徴とする情報検索システム。

【請求項 6】 前記検索結果格納部に格納された文書集合の各文書の特定部分を切り出して格納する文書分類選択格納部を設け、前記自動分類部が、前記文書の特定部分を入力テキストとして文書の自動分類を行うものであり、前記情報空間可視化部が、前記文書の特定部分を入力テキストとして自動分類された情報空間の可視化を行うものであることを特徴とする請求項 5 に記載の情報検索システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 この発明は、文書情報を自動的に分類する文書の自動分類方法、および分類された情報空間を可視化する情報空間の可視化方法、ならびに可視化された情報空間を参照して情報の検索を行う情報検索システムに関し、特に、内容の似た文書が近くに配置されるように 2 次元のセルに配置し、そのセルを代表する語句を表示して、文書情報の情報空間を一覧表示するとともに、文書データを簡単に検索できるようにして、ボトムアップ型の発想支援システムとして機能する情報検索システムに関するものである。

【0002】

【従来の技術】 図 1 3 は従来の情報検索システムの機能構成を示す構成図である。図において、1 は多量の文書情報が格納されている文書データベースであり、2 はこの文書データベース 1 より検索条件に該当した文書データを検索抽出する情報検索部である。また、3 はこの情報検索部 2 にキーワードの論理和や論理積などによる検索条件を入力する検索条件入力部であり、4 は情報検索部 2 より出力される、検索条件に該当した文書の数や、必要に応じて該当する文書の見出しなどの検索結果を出力表示する検索結果表示部である。

【0003】 次に動作について説明する。現在実用化されている特許や科学技術文献などの文書情報を検索する場合、まず、あらかじめ規定されているシソーラスに記載された統制キーワードや、主として文書内に含まれている語句である自由キーワードの論理積や論理和等による検索式を作成し、それを検索条件として検索条件入力部 3 より情報検索部 2 に入力する。情報検索部 2 は文書データベース 1 を検索して、入力された検索条件に該当する文書を抽出し、当該検索条件に合致した文書の数や、必要に応じて検索された文書のタイトルや概要など

の検索結果を検索結果表示部 4 に出力表示する。

【0004】なお、このようなこの発明に関連する従来のベクトルの自動分類方法について記載された文献としては、例えば「プロシーディングス オブ ザ アイ・トリプル・イー (Proceedings of The IEEE)」の第 78 巻第 9 号 (1990 年 9 月) の 1464~1480 ページに掲載された論文「ザ、セルフ・オーガナイズング マップ (The Self-Organizing Map)」などがある。

【0005】

【発明が解決しようとする課題】従来の情報検索システムは以上のように構成されているので、検索条件に該当する文書が文書データベース全体の中でどのような位置を占めるのか明らかでないため、検索された文書が適切なものであるか否かの判断が困難であり、情報検索結果が文書単位でリストとしてしか得られず、内容については順番に本文を参照していく必要があつて、内容の似たものを一括して見るのが困難であるばかりか、文書データベースの全体概要がわからず、文書があらかじめ定められた分類形態を基準に分類されていたとしても、分類のクラスタの相互関係が不明であり、さらに文書データベースにどのような自由キーワードがあるのかわからないなどの問題点があつた。

【0006】この発明は上記のような問題点を解消するためになされたもので、文書を自動分類して内容の近さを反映した 2 次元のセルとして配置し、各セルには分類を代表するキーワードを表示することによって文書データベースの全体構成を把握できるようにし、それを利用して検索キーワードが容易に得られる情報検索システムや、ボトムアップ型の発想支援システムとして機能する情報検索システム、さらには、それに用いられる文書の自動分類方法および情報空間の可視化方法を得ることを目的とする。

【0007】

【課題を解決するための手段】請求項 1 に記載の発明に係る文書の自動分類方法は、文書の語句ベクトルとセルの語句ベクトルの距離を計算して、それが最小となるものをその文書の仮の所属セルとし、その所属セルの語句ベクトルの値をその文書の語句ベクトルの値に近づけるとともに、そのセル近傍のセルの語句ベクトルの値もその文書の語句ベクトルへの近さの度合いに応じて減じて近づける処理を、所定回数もしくは収束するまで繰り返す、その後、各セルの語句ベクトルと文書の語句ベクトルの距離が最小のセルをその文書の所属セルとして、同じセルに所属する文書を内容が類似した文書のクラスタと判断するものである。

【0008】また、請求項 2 に記載の発明に係る文書の自動分類方法は、近傍のセルの語句ベクトルを文書の語句ベクトルに近づけると、語句の文書集合での出現割合に依存して近づける割合を変化させ、文書のクラスタ

の階層性と文書クラスタの所属するセルの領域の包含関係を対応させるものである。

【0009】また、請求項 3 に記載の発明に係る情報空間の可視化方法は、セルに対応する語句ベクトルの値が大きなものから一定数、もしくは所定の閾値以上の語句を、そのセルを代表する語句とし、当該語句の文字列を表示して、上記文書の自動分類方法によって分類された文書集合の情報空間を可視化するものである。

【0010】また、請求項 4 に記載の発明に係る情報空間の可視化方法は、隣接するセルに対応する語句ベクトルの距離に応じて、そのセル間の境界線の属性を変えて表示するものである。

【0011】また、請求項 5 に記載の発明に係る情報検索システムは、情報検索部にて検索された文書集合を格納する検索結果格納部、その文書集合を対象に、請求項 1 または 2 に記載した文書の自動分類方法によって文書の自動分類を行う自動分類部、分類された文書集合の情報空間を、請求項 3 または 4 に記載した情報空間の可視化方法によって視覚化する情報空間可視化部、視覚化された 2 次元の位置を指定することによって特定の文書を選択する文書選択部、および、選択された文書の内容を検索結果格納部より取り出して表示する文書内容表示部を設けたものである。

【0012】また、請求項 6 に記載の発明に係る情報検索システムは、検索結果格納部内の文書の特定部分を切り出して格納する文書分類選択格納部を付加し、当該特定部分を入力テキストとして文書の自動分類および情報空間の可視化を行うものである。

【0013】

【作用】請求項 1 に記載の発明における文書の自動分類方法は、文書の語句ベクトルとセルの語句ベクトルの距離が最小のセルをその文書の仮の所属セルとし、その所属セルの語句ベクトルの値をその文書の語句ベクトルの値に近づけ、またそのセル近傍のセルの語句ベクトルの値もその文書の語句ベクトルへの近さの度合いに応じて近づける学習を行い、学習終了後、各セルの語句ベクトルと文書の語句ベクトルの距離が最小のセルをその文書の所属セルとして、同じセルに所属する文書を内容が類似した文書のクラスタと判断することにより、内容が類似した文書を一括して参照可能とする。

【0014】また、請求項 2 に記載の発明における文書の自動分類方法は、上記学習に際して、語句の文書集合での出現割合によって近づける割合を変化させることにより、文書のクラスタの階層性と文書クラスタの所属するセルの領域の包含関係が対応した文書のクラスタを作成する。

【0015】また、請求項 3 に記載の発明における情報空間の可視化方法は、セルに対応する語句ベクトルの値が大きなものから一定数、もしくは所定の閾値以上の語句をそのセルを代表する語句としてその語句の文字列を

表示することにより、文書集合全体の概要が分かりやすい情報空間の可視化を可能とする。

【0016】また、請求項4に記載の発明における情報空間の可視化方法は、セル間の境界線の属性を、隣接するセルに対応する語句ベクトルの距離に応じて変化させて表示することにより、内容の似たクラスタの領域が分かりやすい情報空間の可視化を可能とする。

【0017】また、請求項5に記載の発明における情報検索システムは、情報検索によって得られた検索結果格納部内の文書集合を対象に、請求項1または2に記載された文書の自動分類方法を用いて文書の自動分類を行い、その自動分類された文書集合の情報空間を、請求項3または4に記載された情報空間の可視化方法によって視覚化し、視覚化された2次元の位置を指定することによって選択した文書の内容を、検索結果格納部より取り出して文書内容表示部に表示することにより、検索された文書集合の全体概要を見ながら個々の文書の内容を確認めることを可能にする。

【0018】また、請求項6に記載の発明における情報検索システムは、検索結果格納部内の文書の特定期間を切り出して文書分類選択格納部に格納し、その特定期間を入力テキストとして文書の自動分類および情報空間の可視化を行うことにより、処理するデータ量を削減して処理時間を短縮する。

【0019】

【実施例】

実施例1. 以下、この発明の一実施例を図について説明する。図1はこの発明による文書の自動分類方法の一実施例における学習フェーズの処理の流れを示すフローチャートであり、図2は同じく分類フェーズの処理の流れを示すフローチャートである。このように、この実施例1による文書の自動分類方法は学習フェーズと分類フェーズとから成っており、以下、まず学習フェーズの動作について説明し、次に分類フェーズの動作について説明する。なお、この明細書中における「語句」という表現は、名詞、動詞などの通常の単語、および句や節など、テキストに含まれる意味のある文字列を表すものである。

【0020】学習フェーズが開始されると、まずステップST100において、文書DOC-1, DOC-2, *40

$$V_{ij} = F_{ij} \times \log(N/N_j) \quad \dots \dots (1)$$

【0023】ただし、上記(1)式において、 F_{ij} は語句word-jが文書DOC-iに出現する頻度、 N は文書集合DOCUMENTSの文書数、 N_j は語句word-jを含む文書の数である。従って、語句word-jが文書集合DOCUMENTSのすべての文書に出現する場合は、 $N_j = N$ となって $\log(N/N_j) = 0$ となるため、 V_{ij} も0となる。これは分類という観点では、すべての文書に出現する語句は、その語句の有無によって文書を分けることができないため、その語

*... , DOC-i, ... , DOC-Nによって構成される文書集合DOCUMENTSに含まれている異なった語句のリストを求めて、それを語句リストWORD-LISTとする。次にステップST110において、ノイズを削減するために、前記語句リストWORD-LISTの語句中より重要なもののみを選んで、それを語句リストWORD-LIST2とする。例えば、出現頻度の高いものは一般的な語句であるため、分類という観点からは重要ではなく、また、出現頻度の低いものは特殊な語句であることが多く、これも分類という観点からは重要ではない。そこで、このステップST110では、語句リストWORD-LISTの各語句が文書集合DOCUMENTS中に含まれる頻度を数え、頻度が第1の閾値FREQUENCY-LOW以下の語句と頻度が第2の閾値FREQUENCY-HIGH以上の語句を語句リストWORD-LISTの語句中より除き、それを語句リストWORD-LIST2とする。なお、このようにして作成された語句リストWORD-LIST2は、語句word-1, word-2, ... , word-i, ... , word-nから構成されているものとする。

【0021】次にステップST120において、2次元に配置されたセルの位置をCELL(x, y)とし、語句リストWORD-LIST2を要素とする語句ベクトルCELL-Vector(x, y)を位置CELL(x, y)のセルに対応するベクトルとする。なお、各語句word-iの初期値は乱数などを使って任意の値にする。ただし、語句ベクトルは単位長に正規化する。次にステップST130に進み、文書集合DOCUMENTSの各文書DOC-iについて、語句リストWORD-LIST2を要素とする語句ベクトルDOC-Vector-iを作成する。語句ベクトルDOC-Vector-iの各語句word-jの値 V_{ij} は、文書に出現する回数が多いほど重要と考えられ、またその語句が出現する文書の数が少ないほど分類という観点からは重要であるので、そのような語句ほど値が大きくなるように、例えば次に示す(1)式によってその値を設定する。

【0022】

句の重要度は0であることを表現している。

【0024】次にステップST140に進んで、後述するステップST141とステップST142の処理を、 $i=1$ から N まで順に T 回繰り返して実行する。なお、その場合、 i は $i=N$ の次は $i=1$ となるものとする。ステップST141では、各文書DOC-iについて、その語句ベクトルDOC-Vector-iと各位置CELL(x, y)のセルの語句ベクトルCELL-Vector(x, y)との距離を計算し、その距離が最小

のものをCELL (p, q) として、その位置のセルをその文書DOC-i が所属する仮のセルとする。次にステップST142において、語句ベクトルDOC-Vector-i をV、語句ベクトルCELL-Vector (x, y) をW (x, y) として、時刻tにおけるその

*のW (x, y) の値をW (x, y) (t) とした時、そのW (x, y) の値を次の(2)式に従って更新する。
【0025】
【数1】

$$= \text{Normalize} [W(x, y)(t) + \alpha(t) \times (V - W(x, y)(t))]$$

$$\alpha(t) = h(t) \times e^{\frac{(x-p)^2 + (y-q)^2}{\delta(t)^2}}$$

$$h(t) = H \times \frac{T-t}{T}$$

$$\delta(t) = \Delta \times \frac{T-t}{T}$$

【0026】ここで、上記(2)式において、Normalize () はベクトルの長さを正規化する関数であり、HおよびΔは定数、α(t) はW (x, y) をVに近づける程度を表す学習係数である。この学習係数α(t) は、時刻tが進むに従ってその大きさh(t) と範囲δ(t) が減少し、t=Tの時刻に0となる。

【0027】ステップST140にて、このステップST141、ステップST142の処理がi=1からNまで順にT回繰り返されるとステップST150に進み、学習フェーズの一連の処理が終了する。

【0028】このステップST150にて学習フェーズが終了すると、次に分類フェーズが開始される。この分類フェーズが開始されると、ステップST160においてまず、位置CELL (x, y) のセルに属する文書の識別子ID-i を保存するためのリストをCELL-Doc (x, y) として、そのリストCELL-Doc (x, y) をnilに初期化する。次にステップST170に進んで、後述するステップST171の処理をi=1からNまで繰り返して実行する。このステップST171では、各文書DOC-i について、その語句ベクトルDOC-Vector-i と各位置CELL (x, y) のセルの語句ベクトルCELL-Vector (x, y) との距離を計算し、それが最小であるセルの位置がCELL (p, q) であった場合に、リストCELL-Doc (p, q) にその文書DOC-i の識別子ID-i を追加する。

【0029】ステップST170にて、このステップST171の処理がi=1からNまで繰り返されるとステップST180に進み、この分類フェーズの一連の処理が終了する。なお、このようにして得られたリストCELL-Doc (x, y) に属する文書が自動分類された文書クラスタである。

【0030】なお、上記実施例1では、最初に与えられた文書集合に属する文書DOC-i を自動的に分類する

..... (2)

ものについて説明したが、学習フェーズが終了した後、未知の文書についてもステップST130と同様の方法でその文書の語句ベクトルを作成し、ステップST171と同様の方法でその文書の属するセルを定めて分類に追加することにより、未知の文書を与えられた文書集合の自動分類と同一の基準で分類することが可能となる。

【0031】さらに、この実施例1では、文書は1つのクラスタに分類されるものとして説明したが、図2のステップST171において、各位置CELL (x, y) のセルの語句ベクトルCELL-Vector (x, y) と各文書DOC-i の語句ベクトルDOC-Vector-i との距離が一定の値以下の位置CELL (x, y) のセルにすべての文書DOC-i が所属するものとして、リストCELL-Doc (x, y) にそれらの文書の識別子ID-i を追加することにより、文書が複数のクラスタに分類されるようにすることも可能である。

【0032】また、図2のステップST171で、各位置CELL (x, y) のセルの語句ベクトルCELL-Vector (x, y) と各文書DOC-i の語句ベクトルDOC-Vector-i との距離が小さいものから一定数の位置CELL (x, y) のセルにすべての文書DOC-i が所属するものとして、リストCELL-Doc (x, y) にそれらの文書の識別子ID-i を追加することによっても、文書が複数のクラスタに分類されるようにすることが可能である。

【0033】また、図2のステップST171で、各位置CELL (x, y) のセルの語句ベクトルCELL-Vector (x, y) と各文書DOC-i の語句ベクトルDOC-Vector-i との距離の分布を計算して、ローカルミニマムとなる位置CELL (x, y) のセルに全ての文書DOC-i が所属するものとして、リストCELL-Doc (x, y) にそれらの文書の識別子ID-i を追加することによっても、文書が複数のク

ラスタに分類されるようにすることが可能である。

【0034】実施例2. 実施例2はこの発明の文書の自動分類方法に関する他の実施例であり、上記実施例1では学習係数が語句の文書集合内での出現の仕方に関係なく一定であったのに対して、語句の文書集合内での出現の仕方に依存して学習係数を変化させている。なお、その学習係数の変化のさせ方については、例えば、ある語句word-iが出現する文書の数 N_i とすると、学習フェーズの初期の段階では N_i が大きな語句の学習係数を、 N_i が小さな語句のそれよりも大きくしておき、学習が進むにつれて N_i の小さな語句の方が N_i の大きな語句よりも学習係数が大きくなるようにする。このように学習させることによって、 N_i の大きな一般的*

$$W(x, y)(t+1)$$

$$= \text{Normalize} [W(x, y)(t) + \alpha(t, N_i) \times (V - W(x, y)(t))]$$

$$\alpha(t, N_i) = h(t, N_i) \times e^{-\frac{(x-p)^2 + (y-q)^2}{\delta(t, N_i)^2}}$$

$$h(t, N_i) = H(N_i) \times \frac{Th(N_i) - t}{Th(N_i)}$$

$$\delta(t, N_i) = \Delta(N_i) \times \frac{T_s(N_i) - t}{T_s(N_i)}$$

$$H(N_i) = H_0 \times [1 - G(N_i)]$$

$$\Delta(N_i) = \Delta_0 \times [1 - G(N_i)]$$

$$Th(N_i) = Th_0 \times G(N_i)$$

$$T_s(N_i) = T_{s0} \times G(N_i)$$

$$G(N_i) = \frac{\log(N/N_i)}{\log N}$$

..... (3)

【0037】なお、上記(3)式において、 H_0 , Δ_0 , Th_0 , T_{s0} は定数であり、 N_i は1から N までの整数値をとる。ここで、関数 $G(x)$ が $G(1) = 1$, $G(N) = 0$ となる単調減少関数であれば、上記

$$G(N_i) = (N - N_i) / (N - 1) \quad \dots\dots\dots (4)$$

【0039】実施例3. 図3はこの発明による情報空間の可視化方法の一実施例における処理の流れを示すフローチャートであり、図4は2次元に配置されたセルの配置例を示す説明図、図5は可視化された情報空間の表示例を示す説明図である。この図4および図5において、5は2次元に配置されたセルであり、図4においてはそれの各々が配置されている位置がCELL(0, 0), CELL(0, 1), ..., CELL(3, 3)で表されている。また、図5において、6はセル5を代表する語句としてそのセル5内に表示された意味のある文字列であり、隣り合ったセル5の間で代表する語句が同一

*な語句の要因を早く学習させることができるようになり、文書のクラスタの階層性と文書クラスタの属するセルの領域の包含関係が対応したものとなる。

【0035】ここで、この実施例2の文書の自動分類方法における学習フェーズおよび分類フェーズでの処理の流れは、図1および図2のフローチャートに示した実施例1の場合と同様である。しかしながら、図1のステップST142にて用いられている $W(x, y)(t)$ の値を更新するための式として、例えば次に示す(3)式を用いている点で異なっている。

【0036】

【数2】

※(3)式における $G(N_i)$ は次の(4)式に示すような他の関数であってもよい。

【0038】

である場合にはその境界線を消去し、それに1つの文字列6を表示している。なお、図4では各セル5が六角形であるものを示したが、4角形など他の形状であってもさしつかえない。

【0040】次にその動作を図3のフローチャートに従って説明する。まず、図1に示した実施例1あるいは実施例2の学習フェーズが終了した後、各位置CELL(x, y)のセル5の語句ベクトルCELL-Vector(x, y)の語句をその値の順にソートする。次にステップST210に進み、その値の大きい順に、あらかじめ定められた数の語句を選択して、それをその位置

CELL (x, y) のセル5を代表する語句とする。次にステップST220において、その選択された語句の文字列6をそれぞれの位置CELL (x, y) のセル5に表示する。

【0041】以下、この文字列6の表示を図5に従って具体的に説明する。図5は国際特許分類のサブクラスG06Fに分類されている特許文書に関して、同一出願人の特許文書41件について自動分類し、その情報空間を可視化した場合の表示例を示したものであり、各位置CELL (x, y) のセル5の語句ベクトルCELL-Vector (x, y) の値が最大の語句を1つだけ選択し、その語句の文字列6を各セル5に表示したものである。なお、この図5においては、表示を見やすくするため、隣接するセル5の相互で代表する語句が同一である場合には、その境界線を消すとともに、その中に文字列6を1つだけ表示するようにしている。例えば、図5の右下のセル5とその左隣のセル5とは代表する語句が同一であるため、両者の間の境界線が消去され、その一方（右下隅のセル5）にのみ共通の文字列6として「処理装置」が表示されている。

【0042】また、この図5では、その右上の部分に「CPU」、「プロセッサ」、「プログラム」などの関係の深い語句の文字列6が表示されたセル5が配置されており、左上の部分には「ディスク装置」、「記憶装置」という関係の深い語句の文字列6が表示されたセル5が配置されている。さらに、その下側には「電力系統」と「知識ベース」の文字列6が表示されたセル5が隣接して配置されているが、これは電力系統の監視に知識ベースを持つエキスパートシステムが利用されていることが推測できる。このように、この実施例3の情報空間の可視化方法によれば、それぞれの代表的な語句の関連が深いセル5が互いに近くなるように配置されて可視化されることとなる。

【0043】なお、この実施例3では、それぞれの位置CELL (x, y) の語句ベクトルCELL-Vector (x, y) の値が最大の語句を1つ選んで、その文字列6を該当するセル5に表示する場合について説明したが、語句ベクトルCELL-Vector (x, y) の値の大きいものから順に一定個数の語句を選択して、その文字列6を表示するようにしても、また、語句ベクトルCELL-Vector (x, y) の値が一定値以上のものをすべて表示するようにしてもよい。なお、その場合、語句ベクトルCELL-Vector (x, y) の値に応じて語句の重要度が区別できるように、文字列6の大きさや書体、さらには表示色などの文字属性を変えるようにしてもよい。

【0044】さらに、各語句word-jの各位置CELL (x, y) のセルでの語句ベクトルCELL-Vector (x, y) の値をグラフとして表示するようにしてもよく、また語句word-jの各位置CELL

(x, y) のセルでの値の分布を計算して、ローカルマキシマムとなるセル5の位置CELL (x, y) にその語句word-jを表示するようにしてもよい。

【0045】実施例4. 図6はこの発明による情報空間の可視化方法の他の実施例における処理の流れを示すフローチャートであり、図7は可視化された情報空間の表示例を示す説明図である。図7において、5はセルであり、7はそのセル5を代表する語句である。また、8は隣接するセル5間の境界線で、セル5を代表する語句ベクトルCELL-Vector (x, y) の距離により、その属性が変えられて表示されるものである。

【0046】次にその動作を図6のフローチャートに従って説明する。まずステップST300において、互いに隣接したセル5をそれぞれセルa、セルbとしたとき、それらの境界をEDGE (a, b) とする。次にステップST310で、すべての境界EDGE (a, b) について、セルaの語句ベクトルCELL-Vector (ax, ay) とセルbの語句ベクトルCELL-Vector (bx, by) の距離を計算する。次にステップST320に進んで、ステップST310で算出された各境界EDGE (a, b) における距離の値を、その最大値のものが1となるように正規化する。次にステップST330で、各境界EDGE (a, b) を示す境界線8の属性値を、その境界EDGE (a, b) の距離の値に従って、あらかじめ定めておいた種類や太さなどを表すものに割り当てる。次にステップST340において、セルaとセルbの境界EDGE (a, b) の境界線8を、その割り当てられた属性によって表示し、ステップST350にて一連の処理を終了する。

【0047】以下、この境界線8の表示を図7を用いて具体的に説明する。ここでは説明を簡単化するため、境界線8の属性の種類は太線と破線の2種類とし、太線は隣り合うセルaの語句ベクトルCELL-Vector (ax, ay) とセルbの語句ベクトルCELL-Vector (bx, by) の距離が大きく、破線はその距離が小さいことを表すものとする。ここで、位置CELL (x, y) のセル5における代表的な語句7をWORDxyとすると、図7は次のことを表していると解釈できる。まず、可視化された情報空間が大きく分けて3つの領域に別れている。すなわち、第1の領域は位置CELL (0, 2)、CELL (1, 2)、CELL (0, 3) およびCELL (1, 3) の4つのセル5による領域である。第2の領域は位置CELL (0, 0)、CELL (1, 0)、CELL (2, 0)、CELL (3, 0)、CELL (0, 1)、CELL (1, 1)、CELL (2, 1)、CELL (2, 2)、CELL (3, 2)、CELL (2, 3) およびCELL (3, 3) の11個のセル5による領域である。第3の領域は位置CELL (3, 1) の1つのセル5による領域である。

【0048】また、第1の領域の各セル5を代表する語

句7であるWORD02、WORD12、WORD03
およびWORD13は互いに連想関係にあり、それぞれのセル5に対応する文書も内容が近い。一方、WORD02とWORD01で代表されるセル5、WORD12とWORD01で代表されるセル5、WORD12とWORD11で代表されるセル5、WORD12とWORD22で代表されるセル5、WORD13とWORD22で代表されるセル5、WORD13とWORD23で代表されるセル5は互いに隣接していても、対応する文書は近い関係にはない。さらに第2の領域内においても、WORD01とWORD10で代表されるセル5、およびWORD11とWORD10で代表されるセル5は近い関係にあるが、WORD01とWORD11で代表されるセル5は隣接していても近い関係にはない。

【0049】実施例5. 図8はこの発明による情報検索システムの一実施例の機能構成を示す構成図である。図において、1は文書データベース、2は情報検索部、3は検索条件入力部、4は検索結果表示部であり、これらは図13に同一符号を付した従来のそれらと同一、もしくは相当部分であるためその説明を省略する。

【0050】また、9は情報検索部2によって検索された文書集合を格納するための検索結果格納部であり、10はこの検索結果格納部9に格納された文書集合を対象にして、請求項1または2に記載された文書の自動分類方法に従って文書の自動分類を行う自動分類部、11は請求項3または4に記載された情報空間の可視化方法に従って、この自動分類部10で自動分類された文書クラスタの代表する語句を2次元に視覚化する情報空間可視化部である。12はこの情報空間可視化部11によって視覚化された2次元の位置を指定することによって、自動分類された文書の中から特定の文書の選択を行う文書選択部であり、13はこの文書選択部12によって選択された文書の内容を検索結果格納部9より取り出して表示する文書内容表示部である。

【0051】次に動作について説明する。ここで、図9はこの実施例5による情報検索システムの処理の流れを示すフローチャートである。まず、ステップST400において、検索条件入力部3より検索条件を入力する。この検索条件はキーワードの論理積や論理和によるものである。次にステップST410において、情報検索部2が文書データベース1を検索してその検索条件に合う文書を抽出し、ステップST420でその検索結果を検索結果表示部4に表示する。なお、この検索結果は通常は検索条件に該当する文書の数であり、必要に応じて文書のタイトルや概要なども表示することがある。次にステップST430に進み、情報検索部2で検索された文書が、内容および数の観点から見て、検索の初期の目的を満たしているか否かを利用者が判断する。その結果、初期の目的を満たしていなければステップST400に戻り、新たな検索条件で再検索を行う。

【0052】一方、初期の目的を満たしている場合には、ステップST440にて検索結果の文書集合を検索結果格納部9に格納する。次にステップST450に進み、検索結果格納部9に格納されている文書集合を対象に、自動分類部10で文書の自動分類を行う。なお、この文書の自動分類は実施例1もしくは実施例2で説明した文書の自動分類方法によって実現される。次にステップST460で情報空間可視化部11によって、自動分類部10が前述のようにして自動分類した情報空間を、代表的な語句に関連の深いものが近くにくるように配置されたキーワードマップの形で可視化表示する。なお、この情報空間の可視化も実施例3もしくは実施例4で説明した情報空間の可視化方法によって実現される。

【0053】次にステップST470において、利用者がこの情報空間可視化部11によって可視化されたキーワードマップを参照して、文書選択部12にて関心のあるセルを選択する。セルが選択されると処理はステップST480に進み、選択されたセルに対応する文書集合がタイトルリストの形で表示される。次にステップST490において、利用者がこのタイトルリストの形で表示された文書集合を参照し、文書選択部12にて関心のある文書を1つ選択する。文書が選択されると処理はステップST500に進み、文書内容表示部13は文書結果格納部9より選択された文書の内容を取り出して表示する。次にステップST510において、利用者が文書内容表示部13に表示された文書の内容を参照し、満足するものであるか否かを判断する。その結果、満足できるものであった場合にはステップST520に進み、一連の処理を終了する。

【0054】一方、満足できるものではなかった場合には、ステップST490に戻って表示されているタイトルリストの中から別の文書を指定してその内容を参照したり、ステップST470に戻ってキーワードマップの別のセルを選択する。このようにして、利用者はキーワードマップで可視化された情報空間を見ながら、満足するまで検索を繰り返す。なお、文書データベース1の大きさが小さい場合には、検索結果格納部9は文書データベース1で代用することも可能である。また、ステップST520での終了は、選択された検索結果の文書集合を対象とした自動分類、情報空間可視化に対するものであり、満足するものがなかった場合にはステップST400に戻り、新たな検索条件を検索条件入力部3に入力する。

【0055】次に、ステップST460からST500までの処理を図について詳細に説明する。図10はこの実施例5による情報検索システムの実行時のスナップショットを示す説明図であり、図中、14は情報空間可視化部11で可視化されたキーワードマップがステップST460において表示されるウィンドウ、15は選択されたセルに対応する文書のタイトルリストがステップS

T480において表示されるウィンドウ、16は選択された文書の内容がステップST500において表示されるウィンドウである。

【0056】ウィンドウ14に表示されたキーワードマップの中の「レジスタ」と表示されているセルを、利用者がマウスなどでポインティングすることによって選択すると、そのセルに対応する自動分類された文書クラスタの文書のタイトルリストがウィンドウ15に表示される。この例では、「マイクロコンピュータ」というタイトルの文書と、「データ処理回路」というタイトルの文書がクラスタになっていたことがわかる。次に、このウィンドウ15上で利用者が「マイクロコンピュータ」の文書を、マウスなどでポインティングすることによって選択すると、その文書の内容がウィンドウ16に表示される。利用者はこのウィンドウ16の表示を参照して、それが満足するものであるか否かを判断する。

【0057】なお、この実施例5では、図9のステップST470で利用者が関心のあるセルを選択して、ステップST480で選択されたセルに対応する文書集合をタイトルリストの形で表示した後、ステップST490で利用者がそのタイトルリストを見て関心のある文書を1つ選択し、ステップST500でその内容を表示するものについて説明したが、ステップST470で利用者が選択したセルに対応する文書集合に含まれている文書の数が1個、または表示画面の制約から決まる所定の個数よりも小さい場合には、ステップST480およびST490を省略して文書の内容を表示するようにしてもよい。

【0058】また、上記実施例5では、文書データベース1の規模が大きく、前処理として検索条件入力部3より入力したキーワードによる検索条件により情報検索を行って、自動分類や情報空間の可視化の対象となる文書の数を絞り込んだ場合について示したが、文書データベース1の規模が小さい場合には、文書データベース1の内容をすべて検索結果格納部9に入れておき、ステップST440からスタートするようにしてもよい。これはステップST400で文書データベース1のすべての文書が該当する検索条件を入力し、ステップST430で「YES」と判断したことに対応する。

【0059】さらに、検索条件入力部3に入力するキーワードの候補となるものを、文書内容表示部13に表示されたテキストの文字列からあらかじめ抽出しておいて、それを選択することにより簡易に入力できるようにしてもよい。

【0060】また、ステップST470で利用者が関心のあるセルを1つまたは複数個選択して、それらのセルに対応する文書集合を検索結果格納部9に格納して、ステップST450に移れるようにしてもよい。

【0061】また、ステップST500で文書の内容を表示するとき、情報空間可視化部11で可視化されたキ

ーワードマップのセルを代表する語句の文字列を、表示色などの属性を変えて分かりやすく表示するようにしてもよい。

【0062】また、検索結果格納部9に格納された文書集合のデータを文書データベース1のデータとして切り替えられるようにしてもよい。

【0063】実施例6. 図11はこの発明による情報検索システムの他の実施例の機能構成を示す構成図で、相当部分には図8と同一符号を付してその説明を省略する。図において、17は情報検索部2にて検索され、検索結果格納部9に格納された文書集合の各文書の特定部分を切り出して格納する文書分類選択格納部である。なお、自動分類部10はこの文書分類選択格納部17に格納されている文書の特定部分を入力テキストとして文書の自動分類を行うものであり、情報空間可視化部11は当該文書の特定部分を入力テキストとして自動分類された情報空間の可視化を行うものである。

【0064】次に動作について説明する。ここで、図12はこの実施例6による情報検索システムの処理の流れを示すフローチャートである。まず、ステップST400からステップST440において、図9に同一のステップ番号を付した実施例5で説明したのと同様の処理が実行される。その後ステップST441に進み、検索結果格納部9に格納されている文書から、あらかじめ定められた特定部分を選択して切り出し、それを文書分類選択格納部17に格納する。

【0065】この特定部分の選択の方法としては、文書の種類に応じて、例えば文書の概要や前書きの第1段落などを選択する。また文書記述のためのISO (International Organization for Standardization: 国際標準化機構) 標準である、SGML (Standard Generalized Markup Language) などの規格に準拠して作成されたタグ付きの文書では、文書の連続した部分だけではなく、文書の連続していない複数の場所から選択することを自動的に行うこともできる。

【0066】次にステップST451において、自動分類部10がこの文書分類選択格納部17に格納されている文書集合を対象に、実施例1もしくは実施例2で説明した文書の自動分類方法による文書の分類が行われる。以下、ステップST460からステップST520において、図9に同一のステップ番号を付した実施例5で説明したのと同様の処理が実行される。このように、この実施例6では、検索結果格納部9の内容よりデータ量かはるかに少ない文書分類選択格納部17の内容を用いて、文書の自動分類および情報空間の可視化が行われることになる。

【0067】なお、上記各実施例では、独立した文書を対象とするものを示したが、文書が互いにリンクで結ば

れたハイパーテキストを対象としてもよく、上記実施例と同様の効果を奏する。

【0068】また、上記各実施例では、1箇所のコンピュータにデータベースとして蓄えられている文書を対象としたものについて説明したが、コンピュータネットワークによって接続された複数のコンピュータに分散して蓄えられた文書を対象にしてもよく、上記実施例と同様の効果を奏する。

【0069】

【発明の効果】以上のように、請求項1に記載の発明によれば、文書の語句ベクトルとセルの語句ベクトルの距離が最小のセルをその文書の仮の所属セルとして、その所属セルの語句ベクトルの値をその文書の語句ベクトルの値に近付けるとともに、そのセル近傍のセルの語句ベクトルの値もその文書の語句ベクトルへの近さの度合いに応じて近付ける学習を行って、その学習の終了後に、各セルの語句ベクトルと文書の語句ベクトルの距離が最小のセルをその文書の所属セルとして、同じセルに所属する文書を内容が類似した文書のクラスと判断するように構成したので、文書を語句のベクトルとして表現し、そのベクトル表現された文書を自動分類することが可能となり、内容の類似した文書を一括して見ることができる文書の自動分類方法が得られる効果がある。

【0070】また、請求項2に記載の発明によれば、学習の際に語句の文書集合における出現割合に基づいて近付ける割合を変化させるように構成したので、語句の文書集合での出現の分布に依存した学習を行わせることが可能となり、文書のクラスターの階層性と文書クラスターの属するセルの領域の包含関係が対応した文書のクラスターを作成できる効果がある。

【0071】また、請求項3に記載の発明によれば、自動分類された情報空間について、セルに対応する語句ベクトルの値が大きなものから一定数、もしくは所定の閾値以上の語句をそのセルを代表する語句としてその文字列を表示するように構成したので、自動分類された文書集合を代表する語句に関連の深いセルが近くに配置されるように表示して、文書集合の情報空間を可視化することが可能となつて、文書集合全体の概要が分かりやすい情報空間の可視化方法を得ることができ、さらに、どのようなキーワードがあるかを容易に知ることが可能となるばかりか、分類されたクラスターの相互関係が把握しやすくなる効果がある。

【0072】また、請求項4に記載の発明によれば、隣接するセルに対応する語句ベクトルの距離に応じて、セル間の境界線の属性を変化させて表示するように構成したので、隣り合ったセルの類似度を判断することが容易となつて、文書集合全体の概要がより分かりやすいものとなり、また内容の似たクラスターの領域が分かりやすくなる効果がある。

【0073】また、請求項5に記載の発明によれば、情

報検索によって検索結果格納部に格納された文書集合を対象に、請求項1または2に記載した文書の自動分類方法による文書の自動分類を行い、自動分類された文書集合の情報空間を、請求項3または4に記載した情報空間の可視化方法で視覚化し、視覚化された2次元の位置指定によって選択した文書の内容を検索結果格納部より取り出して表示するように構成したので、可視化された情報空間のセルを指定してそのセルに属する文書のリストを表示し、さらに表示されたリストの文書を指定することによって所望の文書の内容を表示することが可能となり、検索された文書集合の全体概要を見ながら個々の文書の内容を確かめることができる情報検索システムが得られる効果がある。

【0074】また、請求項6に記載の発明によれば、検索結果格納部内の文書の特定部分を切り出して文書分類選択格納部に格納しておき、文書の自動分類および情報空間の可視化をその文書の特定部分を入力テキストとして行うように構成したので、検索結果格納部よりはるかにデータ量の少ない文書部分選択格納部の内容を用いて、文書の自動分類および情報空間の可視化を行うことが可能となり、処理するデータ量が減ることによって処理が高速化される効果がある。

【図面の簡単な説明】

【図1】 この発明の実施例1による文書の自動分類方法の学習フェーズの処理の流れを示すフローチャートである。

【図2】 上記実施例における分類フェーズの処理の流れを示すフローチャートである。

【図3】 この発明の実施例3による情報空間の可視化方法の処理の流れを示すフローチャートである。

【図4】 上記実施例におけるセルの配置例を示す説明図である。

【図5】 上記実施例における可視化された情報空間の表示例を示す説明図である。

【図6】 この発明の実施例4による情報空間の可視化方法の処理の流れを示すフローチャートである。

【図7】 上記実施例における可視化された情報空間の表示例を示す説明図である。

【図8】 この発明の実施例5による情報検索システムの機能構成を示す構成図である。

【図9】 上記実施例の処理の流れを示すフローチャートである。

【図10】 上記実施例における実行時のスナップショットを示す説明図である。

【図11】 この発明の実施例6による情報検索システムの機能構成を示す構成図である。

【図12】 上記実施例の処理の流れを示すフローチャートである。

【図13】 従来の情報検索システムの機能構成を示す構成図である。

10

20

30

40

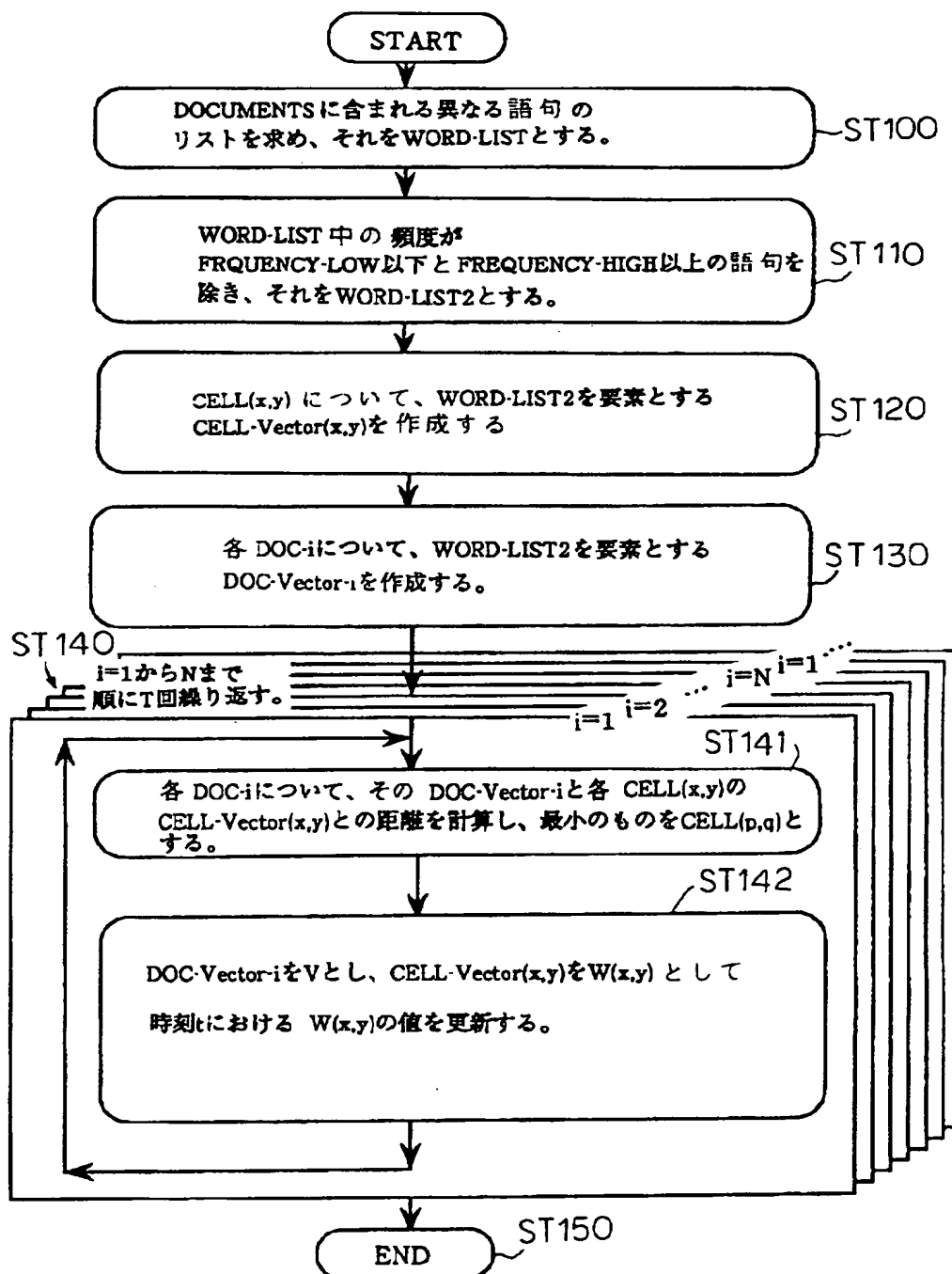
50

【符号の説明】

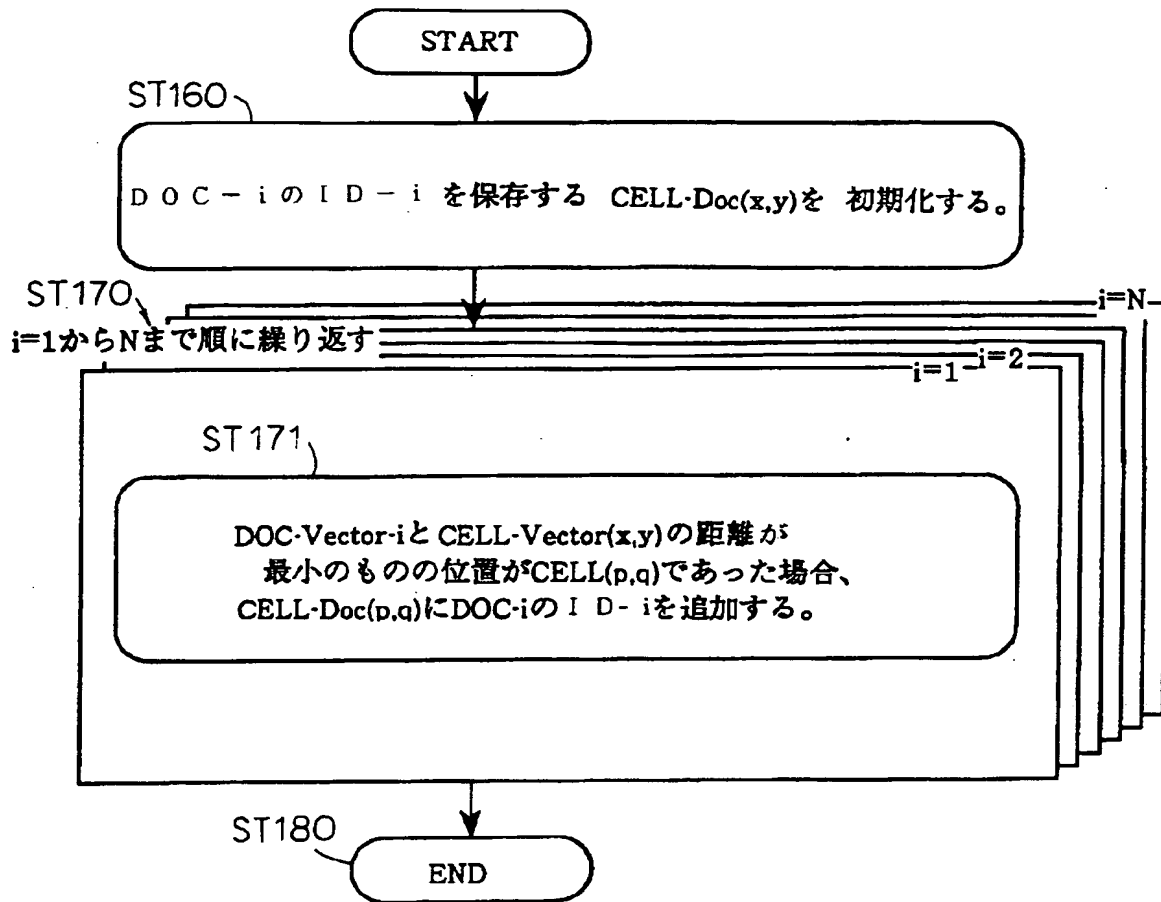
1 文書データベース、2 情報検索部、3 検索条件
入力部、4 検索結果表示部、5 セル、6 文字列、*

* 8 境界線、9 検索結果格納部、10 自動分類部、
11 情報空間可視化部、12 文書選択部、13 文
書内容表示部、17 文書分類選択格納部。

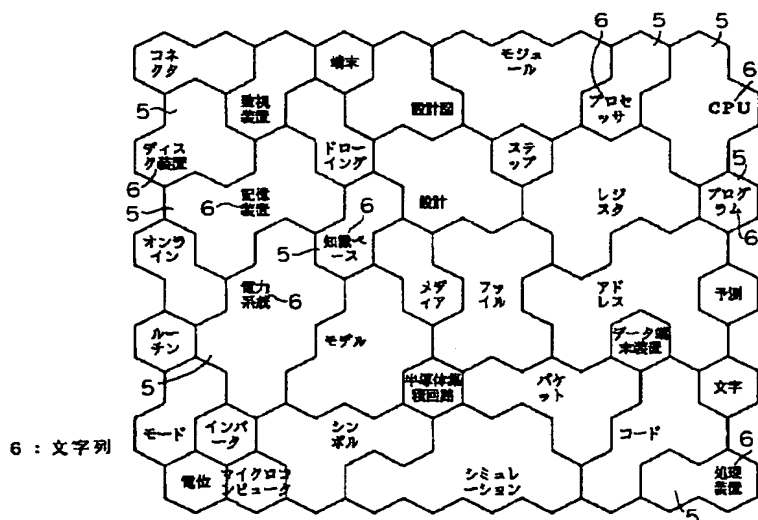
【図1】



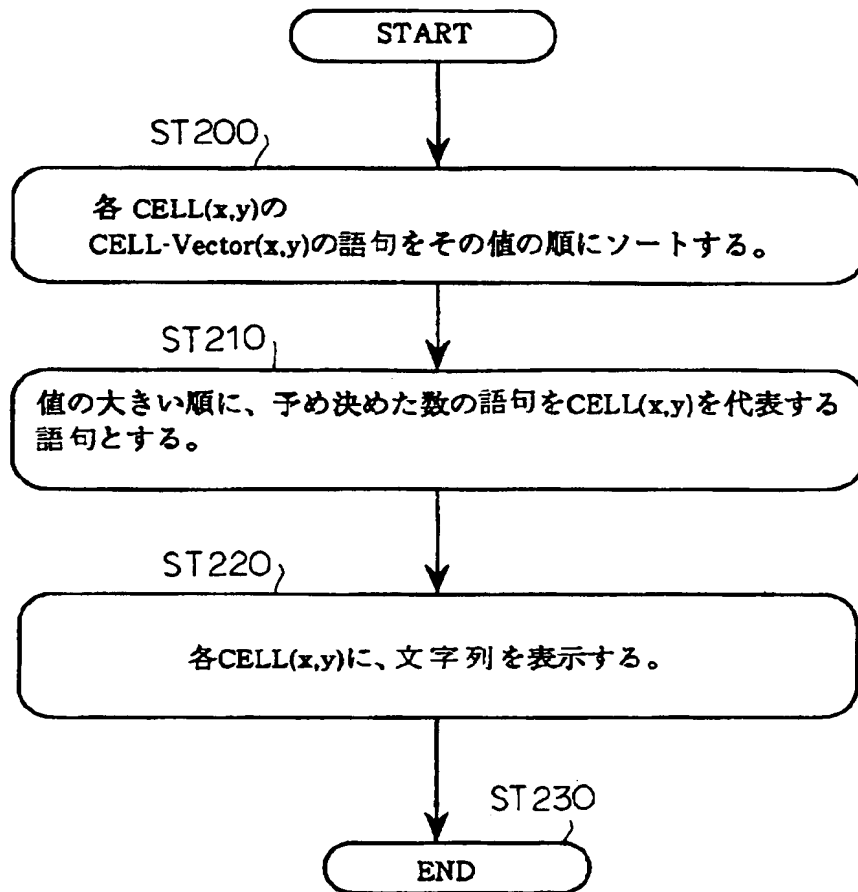
【圖 2】



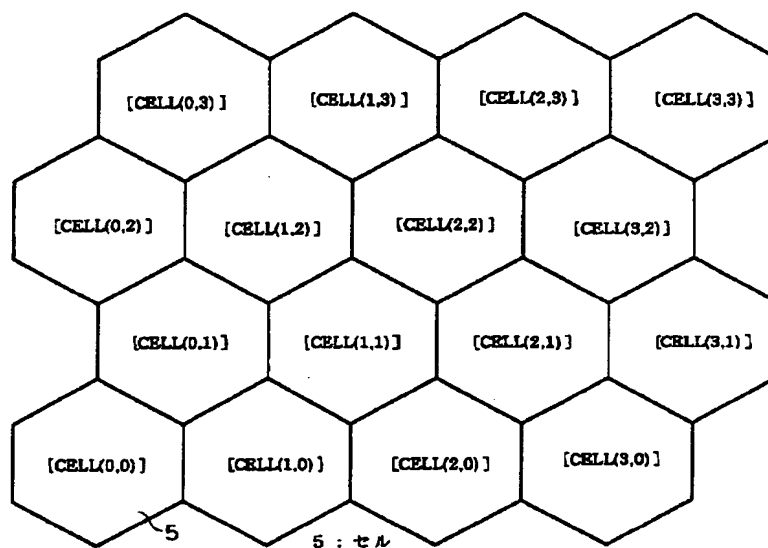
【図 5】



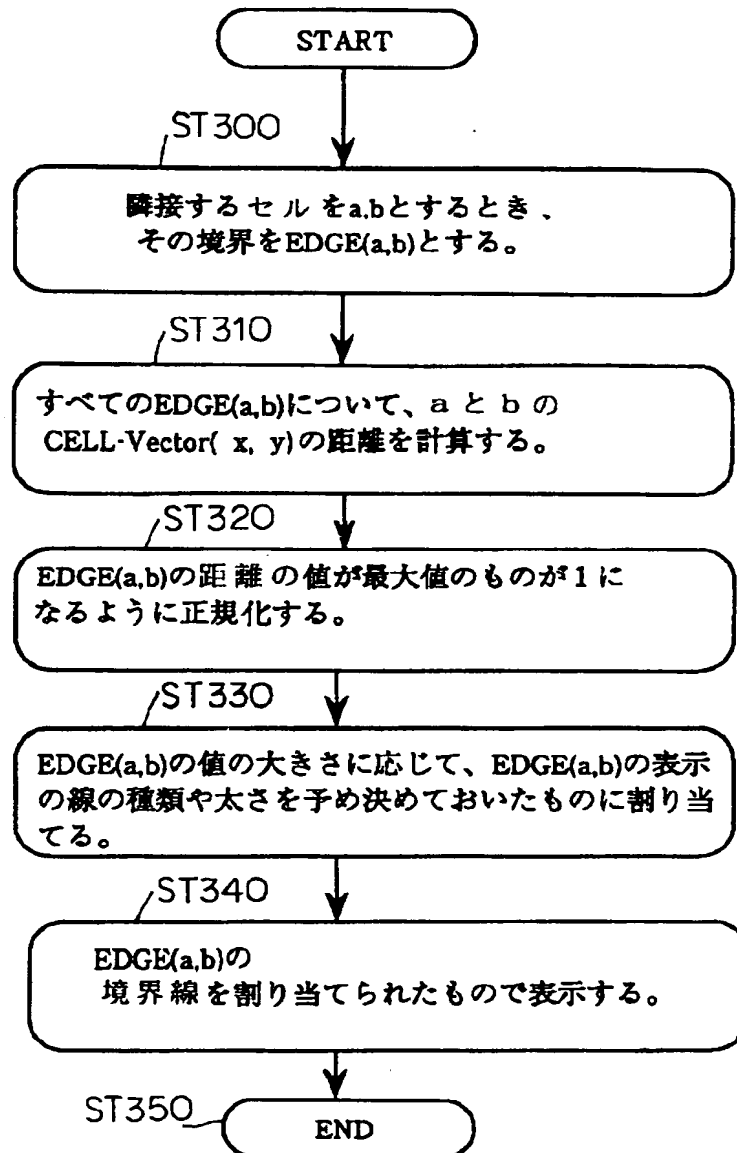
【図 3】



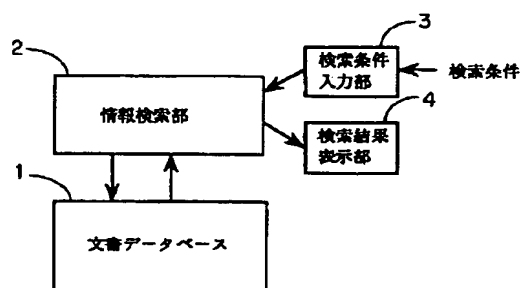
【図 4】



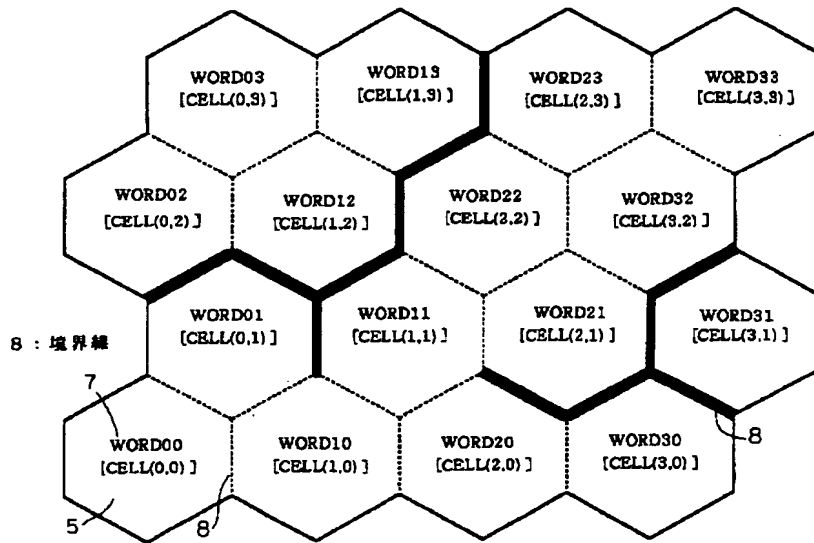
【図6】



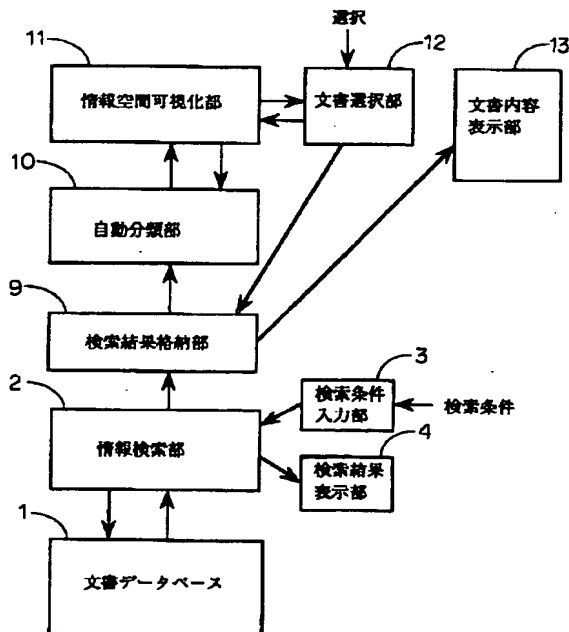
【図13】



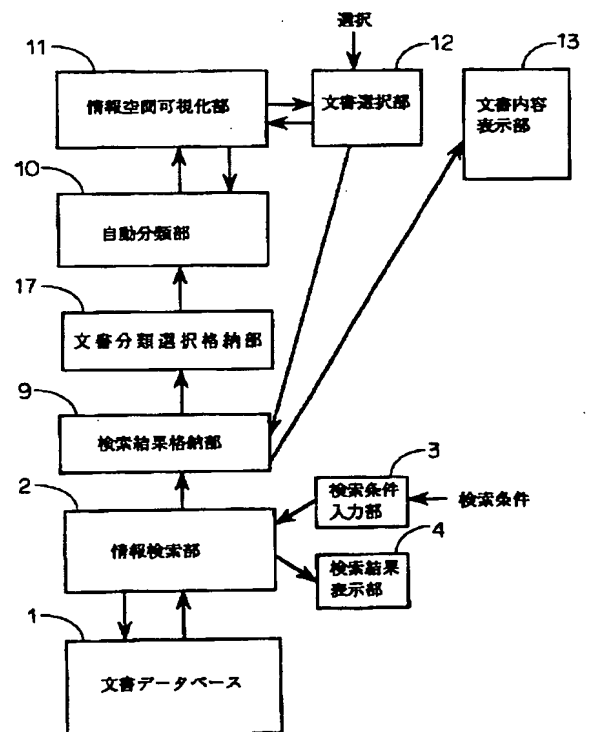
【図 7】



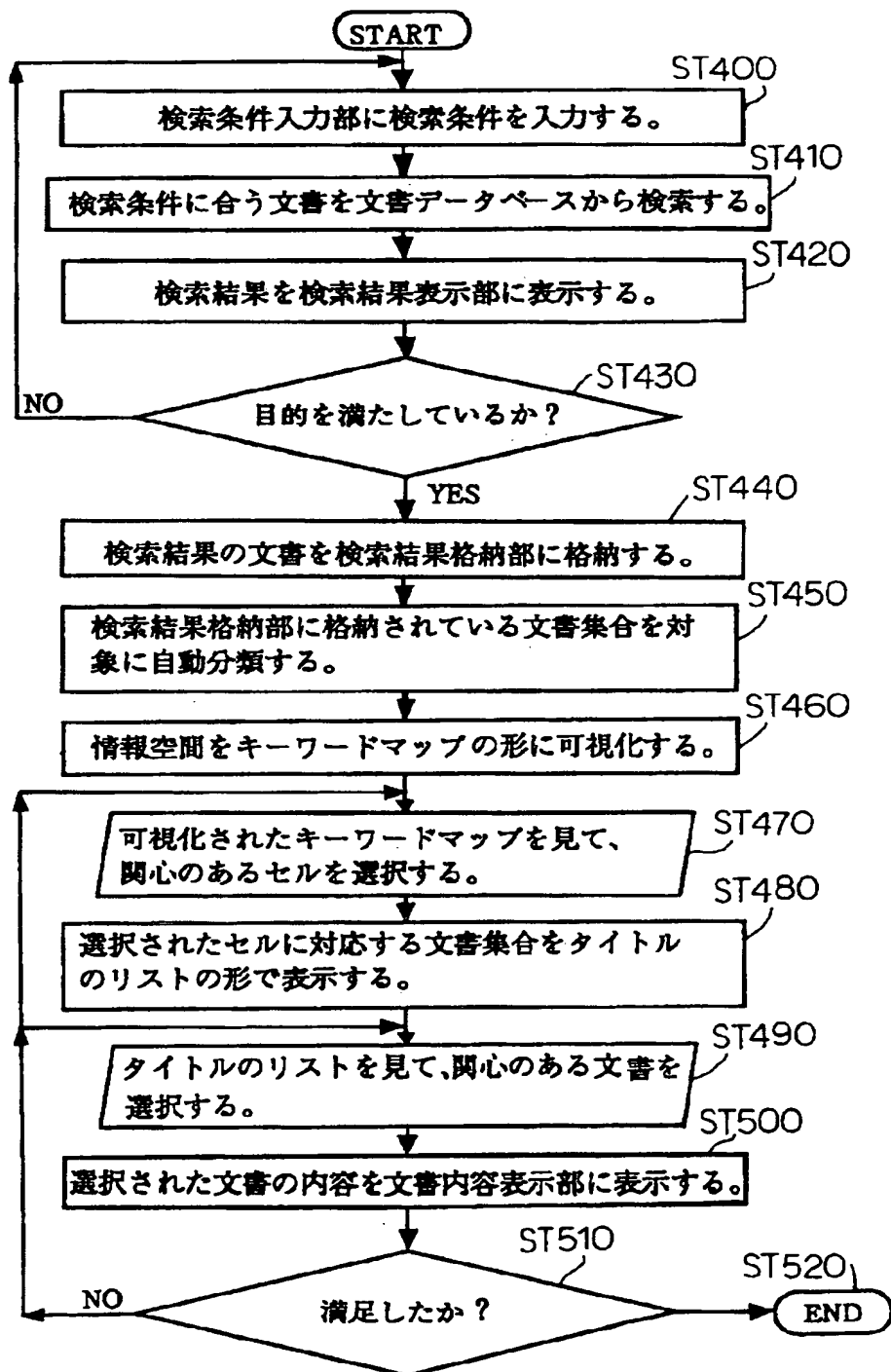
【図 8】



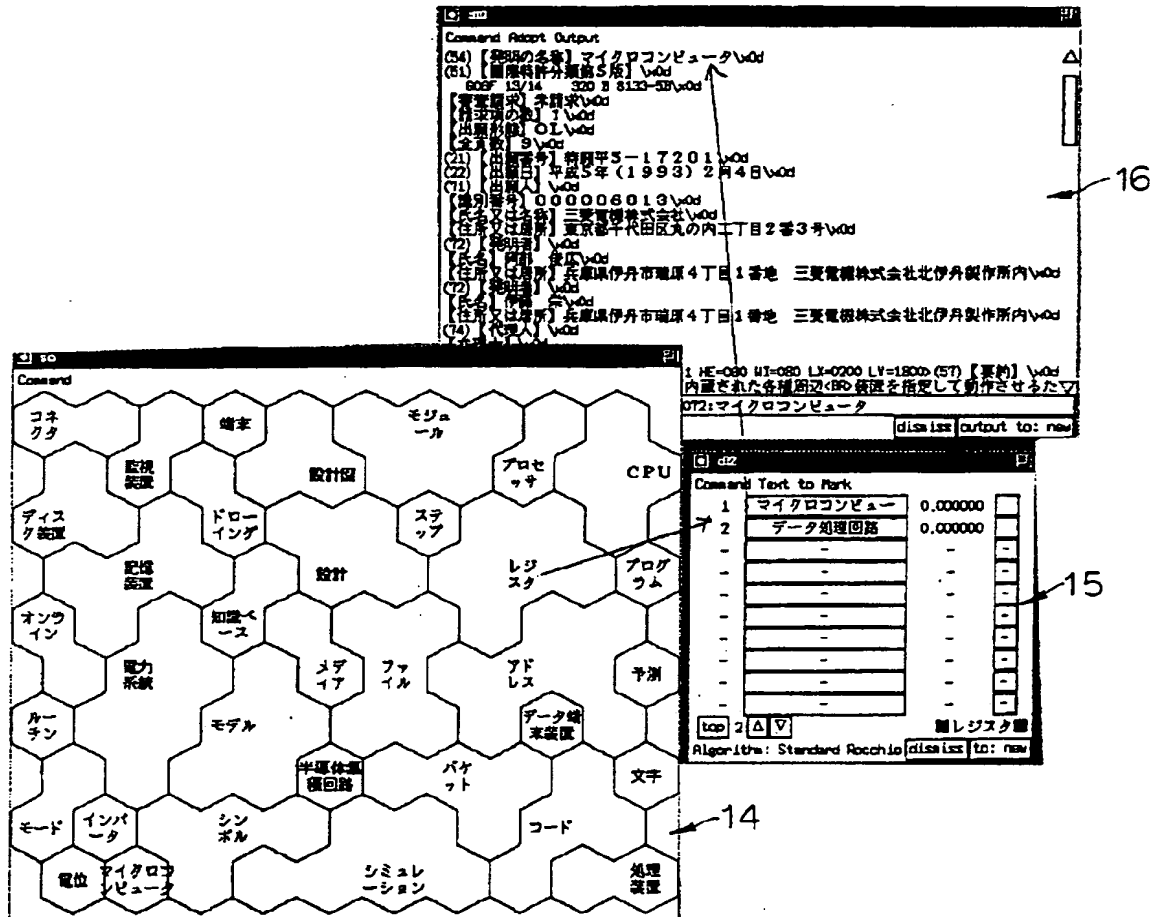
【図 11】



【図 9】



【図10】



【図12】

